Amid genomic data explosion, scientists find proliferating errors

April 30 2019

Washington State University researchers found a troubling number of errors in publicly available genomic data as they conducted a large-scale analysis of protein sequences.

The work, published in the journal *Frontiers in Microbiology*, the world's most cited microbiology journal, could have important implications for future genomic research.

The interdisciplinary team of scientists initially set out to find evidence of a minimal set of proteins that a Proteobacteria needs for survival. Their dataset consisted of nearly nine million protein sequences clustered by similarity from more than 2,300 bacterial genomes.

A genome is the complete set of genes in a cell or organism, and the genes provide instructions for building the proteins that make up all organisms.

As they searched through the massive dataset for four specific proteins thought to be part of a minimal genome for Proteobacteria, they discovered that only one of the four proteins they were looking for was shared by all the bacteria. They also found large numbers of errors in the publicly available data.

"We found that for each of the proteins, there were mistakes in annotation of their genes, which resulted in truncated or missing sequences," said Shira Broschat, a professor in the School of Electrical

Engineering and Computer Science.

The immense amounts of data being created by next-generation sequencing technologies make the kind of annotation errors the WSU team found especially problematic, said Svetlana Lockwood, lead author on the paper and a Ph.D. graduate in computer science from WSU.

"A single annotation error can propagate rapidly because scientists build on previous annotation when they sequence new genomes," she said.

While it took 13 years and $2.7 billion to sequence the human genome as part of the Human Genome Project in 2003, that same work can now be done in a single hour for less than $1500.

"Just in the last two years, researchers have sequenced more than twice the number of bacterial genomes as they did in the twenty years before that," Broschat said.

While this isn't the first paper to note the existence of annotation errors, the WSU team's work lists and explains the various kinds of annotation errors that are currently found in the genomic sequencing data.

"With the scale of mis-annotation we found, researchers have to reevaluate the reliability of publicly available genome data for use in big data applications," Broschat said.

According to Kelly Brayton, a professor in the Department of Veterinary Microbiology and Pathology, the errors are due to human and technological factors. Errors often happen because of imperfect DNA sequencing technology, which provides the information on the base pairs in DNA segments. They can also occur due to confusion and lack of knowledge about the proteins as well.

The team used state-of-the-art software and a high performance computing cluster on the PNNL campus to work on their dataset, the largest of its kind analyzed to date. The data was collected from a database provided by the National Center for Biotechnology Information, part of the United States National Library of Medicine, the world's largest medical library, and the work was funded by the National Science Foundation.

Broschat and Brayton are now working on a tool to find annotation errors in biological datasets, which would be of great use to anyone working in the life sciences.

More information: Svetlana Lockwood et al, Whole Proteome Clustering of 2,307 Proteobacterial Genomes Reveals Conserved Proteins and Significant Annotation Issues, *Frontiers in Microbiology* (2019). DOI: 10.3389/fmicb.2019.00383